Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media

GEORGIOS PALTOGLOU
University of Wolverhampton

and
MIKE THELWALL
University of Wolverhampton

Sentiment analysis is a growing area of research with significant applications in both industry and academia. Most of the proposed solutions are centered around supervised, machine learning approaches and review-oriented datasets. In this paper, we focus on the more common informal textual communication on the web, such as online discussions, tweets and social network comments and propose an intuitive, less domain-specific, unsupervised, lexicon-based approach that estimates the level of emotional intensity contained in text in order to make a prediction. Our approach can be applied to, and is tested in, two different but complementary contexts: subjectivity detection and polarity classification. Extensive experiments were carried on three real-world datasets, extracted from online social websites and annotated by human evaluators, against state-of-the-art supervised approaches. The results demonstrate that the proposed algorithm, even though unsupervised, outperforms machine learning solutions in the majority of cases, overall presenting a very robust and reliable solution for sentiment analysis of informal communication on the web.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms: Languages, Experimentation, Algorithms

Additional Key Words and Phrases: Opinion Mining, Sentiment Analysis, Social Media

1. INTRODUCTION

The proliferation of Facebook, MySpace, Twitter, blogs, forums and other online means of communication has created a digital landscape where people are able to socialize and express themselves through a variety of means and applications. Indicative of this new trend of social interaction is the fact that popular social websites are challenging the popularity of established search engines [Thelwall 2008; Harvey 2010] as the most visited websites on the web.

Sentiment analysis, the process of automatically detecting if a text segment con-

Author's address: School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, UK

tains emotional or opinionated content and determining its polarity (e.g. 'thumbs up' or 'thumbs down'), is a field of research that has received significant attention in recent years, both in academia and in industry. One of the main reasons for this phenomenon is the aforementioned increase of user-generated content on the web which has resulted in a wealth of information that is potentially of vital importance to institutions and companies [Wright 2009]. As a result, most research has centered around product reviews [Pang and Lee 2008; Dave et al. 2003; Turney 2002], aiming to predict whether a reviewer recommends a product or not, based on the textual content of the review.

The focus of this paper is different: the far more widespread informal, social interactions on the web. In this context, sentiment analysis aims to detect whether a textual communication contains expressions of *private states* [Quirk 1985] and subsequently whether it expresses a positive (e.g. excitement, enthusiasm) or negative (e.g. argument, irony, disagreement) emotion. The unprecedented popularity of social platforms such as Facebook, Twitter, MySpace and others has resulted in an unparalleled increase of public textual exchanges that remain relatively unexplored, especially in terms of their emotional content.

One of the main reasons for this discrepancy is the fact that no clear "golden standard" exists in the domain of informal communication, in comparison to the more *de facto* domain of movie or general product reviews. In the latter case, it is generally easy to detect whether a review, extracted from a website is positively or negatively oriented towards a specific product, by simply extracting the specific metadata that accompanies the review, such as the "number of stars" or the "thumbs up/down". As a result, there is an abundance of product related sentiment analysis datasets [Pang et al. 2002; Blitzer et al. 2007; Baccianella et al. 2010], but only a few for the more general field of informal communication on the web [Pak and Paroubek 2010; Mishne 2005].

The differences between the two domains are numerous. First, reviews tend to be longer and more verbose than typical online social interactions, which may only be a few words long [Thelwall and Wilkinson 2010]. Second, social exchanges on the web tend to be much more diverse in terms of their topics with issues ranging from politics and recent news to religion. In contrast, product reviews by definition have a specific subject, i.e. the product under discussion. Last, informal communication often contains numerous, non-standard spellings [Thelwall and Wilkinson 2010], resulting in a very heterogeneous content. For example, Thelwall [2009] reports that 95% of the exchanged comments in MySpace contain at least one abbreviation (such as "m8" for "mate") of standard English.

The above reasons (i.e. abundance of training documents, easily extractable classes, long documents) have channeled the research in the field towards review-related applications, creating a suitable environment for machine learning approaches, while undermining the utilization of unsupervised approaches, which, although perhaps not particularly effective for product related datasets, may demonstrate significant advantages over machine learning approaches in socially-driven environments, where training data (i.e. documents with pre-assigned classes) is difficult to come by and often requires extensive human labor.

In this paper, we propose an unsupervised, lexicon-based classifier that estimates

the level of emotional valence in text in order to make a prediction, explicitly designed to address the issue of sentiment analysis in such environments. We have added an extensive list of linguistically-driven functionalities to the classifier, such as: negation/capitalization detection, intensifier/diminisher detection and emoticon/exclamation detection, all of which contribute to the final prediction. The proposed algorithm is applicable in two different but complementary settings: opinion detection (i.e. detecting whether the text contains an expression of opinion or is objective) and polarity detection (i.e. predicting whether a subjective text is negatively or positively oriented), overall constituting a solution that can be applied without any modification or training to a wide set of environments and settings. Extensive experiments presented in section 5, using real-world datasets from social websites and annotated by human assessors, demonstrate that the lexicon-based approach performs surprising well in the vast majority of cases, persistently outperforming state-of-the-art machine learning solutions, even in environments where there is a significant number of training instances for the latter.

Although extensive research has been directed towards automatically creating sentiment lexicons [Takamura et al. 2005; Baccianella et al. 2010; Hassan and Radev 2010], these attempts typically test the produced dictionaries against a "gold-standard" lexicon (e.g. General Inquirer [Stone et al. 1966]), stopping short of examining their effectiveness in classification scenarios. An important contribution of the present work is to demonstrate how such sentiment lexicons (created automatically or not) accompanied by intuitive syntactic and prose rules can be effectively utilized for this task. Although previous work has also studied the effect of negation or valence shifters detection, e.g. Kennedy and Inkpen [2006], this is the first time that an extensive list of stylistic and linguistic modifiers are utilized. Additionally, although the proposed method may be characterized as self-evident, there is a lack of similar approaches in the field to the best of our knowledge. Lastly, it must be noted that this is the first time that the effectiveness of a extensive lexicon-based solution is compared against standard state-of-the-art machine-learning classifiers for the tasks that are being examined here, in multiple test collections and is found to outperform supervised methods, demonstrating that in certain environments simple and intuitive methods are able to offer a robust and adequate solution.

The rest of the paper is structured as follows. The next section provides a brief overview of relevant work in sentiment analysis. Section 3 presents the lexicon-based classifier and section 4 describes the data sets that are utilized and details the experimental setup while section 5 presents and analyzes the results. Finally, we conclude and present some potential future directions of research in section 6.

2. PRIOR WORK

Sentiment analysis has been a popular research topic in recent years. Most research has focused on analyzing the content of either movie or general product reviews (e.g. [Pang et al. 2002; Blitzer et al. 2007; Dave et al. 2003]), but attempts to expand the application of sentiment analysis to other domains, such as political debates [Thomas et al. 2006; Lin et al. 2006] and news [Devitt and Ahmad 2007] have also been developed. Pang and Lee [2008] present a thorough analysis of the work in the field. In this section we will focus on the more relevant to our work approaches.

2.1 Machine-learning approaches

Pang et al. [2002] were amongst of the first to explore the sentiment analysis of reviews focusing on machine-learning approaches. They experimented with three different algorithms: Support Vector Machines (SVMs), Naive Bayes and Maximum Entropy classifiers, using a variety of features, such as unigrams and bigrams, part-of-speech tags and binary and term frequency feature weights. Their best accuracy attained in a dataset consisting of movie reviews used a SVM classifier with binary features, although all three classifiers gave very similar performance.

Later, the same authors presented an approach based on detecting and removing the objective parts of documents [Pang and Lee 2004]. The results showed an improvement over the baseline of using the whole text using a Naive Bayes classifier but only a slight increase compared to using a SVM classifier on the entire document.

Most other approaches in the field have focused on extending the feature set with semantically or linguistically-driven features in order to improve classification accuracy. For example, Mullen and Collier [2004] used SVMs and enhanced the feature set with information from a variety of diverse sources, such as Osgood's Theory of Semantic Differentiation [Osgood 1967] and Turney's semantic orientation [Turney 2002] resulting in an improvement over the baseline of using only unigrams. Similarly, Whitelaw et al. [2005] used fine-grained semantic distinctions in the feature set to improve classification. Their approach was based on a semi-automatically created dictionary of adjectives with their respective appraisal attribute values, which resulted in 1329 adjectives and modifiers in several taxonomies of appraisal attributes. Conjunctions of the produced lexical lemma with different appraisal groups and bag-of-word approaches were used as features to a Support Vector Machine classifier.

Wilson et al. [2009] studied the effect of analyzing the context of words with a known prior polarity. They hypothesized that it may have an impact on how the words are used in the text, e.g. a positive word may have its semantic orientation negated, and concluded that for improved effectiveness it is important to distinguish when a polar term is used in a neutral context. Zaidan et al. [2007] utilized additional human annotation in order to improve classification accuracy, introducing the notion of "annotator rationales", i.e. manually extracted words or phrases that strongly indicate the polarity of a document. Very recently, Yessenalina et al. [2010] explored methods of automatically detect such phrases, using OpinionFinder [Wilson et al. 2005] and a polarity lexicon [Wilson et al. 2005a].

2.2 Dictionary-based approaches

Dictionary/lexicon-based sentiment analysis is typically based on lists of words with pre-determined emotional weight. Examples of such dictionaries include the General Inquirer (GI) [Wilson et al. 2005b] and the "Linguistic Inquiry and Word Count" (LIWC) software [Pennebaker J. and R. 2001], which is also used in the present study. Both lexicons are built with the aid of "experts" that classify tokens in terms of their affective content (e.g. positive or negative). The "Affective Norms for English Words" (ANEW) lexicon [Bradley and Lang 1999] contains ratings of terms on a nine-point scale in three individual dimensions: valence, arousal and

dominance. The ratings were produced manually by psychology students. Ways to produce such "emotional" dictionaries in an automatic or semi-automatic fashion have also been introduced in research [Turney and Littman 2002; Brooke et al. 2009; Baccianella et al. 2010; Hassan and Radev 2010]. Dictionary-based approaches have been utilized in psychology or sociology oriented research [Chung and Pennebaker 2007; Slatcher et al. 2007].

Turney's PMI-IR algorithm [Turney 2002] is one of the few unsupervised approaches that has been presented in the field of sentiment analysis. It is based on automatically estimating the semantic orientation of phrases extracted from documents. The orientation of the extracted phrases is estimated based on their collocation with certain preselected reference words, using a search engine as a *reference corpus*[1].

Qiu et al. [2009] presented a hybrid solution that combines both approaches. The idea is based on an iterative process that uses a dictionary in order to initially classify documents and then train a machine-learning classifier. The trained classifier is then used in order to revise the results of the first stage classification and then it is re-trained on the refined classifications. The process is continued until the classification results over two consecutive iterations are the same.

Our approach shares some common features with the Opinion Observer system by Ding et al. [2008], in that it also uses an opinion lexicon to derive the polarity of a set of terms and it incorporates a *negation detection* module that detects when the semantic orientation of a opinion word is reversed. Nonetheless, the approach presented here determines the overall polarity of a document based on the intensity of its emotional content (e.g. a document may be "very positive" and at the same time "mildly negative", such as "I love you, I miss you!!"). Additionally, Opinion Observer attempts to extract the semantic orientation of ambiguous words based on their collocation in other reviews of the same product, a process that is inapplicable in the setting of social interactions that we are examining. In addition, our approach incorporates a much wider set of linguistic and stylistic detectors, in order to better capture the intensity of expressed emotion, such as capitalization,emoticons etc.

Lastly, some similarities are also present between our approach and the Affect Analysis Model (AAM) by Neviarouskaya et al. [2007] and the SentiStrength algorithm by Thelwall et al. [2010], in that they also use emotional lexicons and incorporate modifiers, such as "very","hardly", "not" etc., although those are utilized differently in our approach. However, in contrast to the present study their aim is to predict the emotional intensity and strength of textual utterances, the former in several different categories, e.g. anger, disgust, etc. and the latter in terms of positivity and negativity. Additionally, AAM utilizes a syntactic parser, whose effectiveness is doubtful in the domain of informal, social communication where, as seen above, syntactic and orthographic errors are the norm and its effectiveness was tested on a limited number of manually selected 160 sentences providing no comparison between the proposed system and machine-learning approaches. In contrast, in the present work we report experiments on three different and diverse real-world datasets, some of which contain several thousand posts and compare the

[1]In this aspect, it is debatable whether the approach is truly unsupervised, but it is mostly considered as such in research.

effectiveness of our solution against state-of-the-art supervised algorithms.

It is important here to note that most of the previous approaches focus only on one of two problems: subjectivity or polarity detection. In comparison, our proposed algorithm is applicable in both contexts and can therefore offer a much more robust solution to the general problem of detecting emotional content on the web.

3. LEXICON-BASED CLASSIFICATION

The proposed classifier[2] is a typical example of an *unsupervised* approach, because it can function without any *reference corpus* and doesn't require any training (i.e. can be applied "off-the-shelf"). The classifier is based on estimating the intensity of negative and positive emotion in text in order to make a ternary prediction for subjectivity and polarity, i.e. the output of the classifier is one of $\{0, +1, -1\}$. The notion that both negative and positive emotion is present in a text may seem somewhat peculiar, but is in accordance with a number or psychological studies (i.e. [Schimmack 2001; Cornelius 1996; Fox 2008]) and is therefore adopted as the underlying premise of our approach. The level of valence in each scale is measured in two independent ratings $\{C_{pos}, C_{neg}\}$; one for the positive dimension ($C_{pos} = \{1, 2 \ldots, 5\}$) and one for the negative ($C_{neg} = \{-1, \ldots, -5\}$), where higher absolute values indicate stronger emotion and values $\{1, -1\}$ indicate lack of (i.e. objective text).

For example, a score like $\{+3, -1\}$ would indicate the presence of only positive emotion, $\{+1, -4\}$ would indicate the presence of (quite strong) negative emotion and $\{+4, -5\}$ would indicate the presence of both negative and positive emotion. In order to make a ternary prediction, the most prevalent emotion, i.e. the one with the highest absolute value, is returned as the final judgement, e.g. positive in the first example above and negative in the other two. For example, the sentence "I hate the fact that I missed the bus, but at least I am glad I made it on time:-)" expresses both negative and positive emotion, where the latter is considered dominant. We solve conflicts of equality (e.g. $\{+3, -3\}$) by taking into consideration the number of positive and negative tokens and giving preference to the class with the largest number of tokens. A document is classified as objective if its scores are $\{+1, -1\}$. Note that the $\{C_{pos}, C_{neg}\}$ ratings are only used as an intermediate step in making the final prediction.

The algorithm is based on the *emotional* dictionary from the "Linguistic Inquiry and Word Count" (LIWC) software[3] [Pennebaker J. and R. 2001] which was derived from a number of psychological studies and maintains an extensive dictionary list along with human assigned emotional categories for each lemma. We use the weights assigned to the LIWC lexicon by Thelwall et al. [2010]. The justification of utilizing the particular word-list is twofold: a) it has been extensively used in psychology-related research and its development is solely driven by psychological studies, in comparison to other review-oriented, e.g. [Turney and Littman 2002], or linguistic-based, e.g. [Baccianella et al. 2010], dictionaries, and b) it is better suited for

[2]The classifier is publicly available for research purposes, both as a *.jar* file with an open API and a command-line interface, as well as a C++ *.dll* at `http://www.CyberEmotions.eu`.
[3]`http://www.liwc.net`

the type of informal communication [Thelwall et al. 2010] usually found in social media. All dictionary lemmas (as well as processed text) are stemmed using the porter stemmer.

Given a document d, the algorithm detects all words that belong to the emotional dictionary and extracts their polarity and intensity. We modify the initial term scores with additional, prose-driven functionalities such as: *negation detection* (e.g. "good" versus "not good"), *capitalization detection* (e.g. "bad" versus "BAD"), *exclamation and emoticon detection* (e.g. "happy!!" and ":-)") *intensifiers* (e.g. "liked" versus "liked very much") and *diminishers* (e.g. "excellent" versus "rather excellent"), to produce the final document scores. The list of modifiers and their respective weights was adapted from Neviarouskaya et al. [2007] and Thelwall et al. [2010] and manually checked for duplicates and conflicts.

The modules function in the following way: the neighborhood of every word that is present in the text and belongs to the LIWC lexicon is scanned for "special" terms, such as negators (e.g. "not") intensifiers (e.g. "very") or diminishers (e.g. "little"). In order to capture long-distance phenomena, e.g. "I don't think this is a good movie...", neighborhood is defined as the area 5 words before and after the emotional term or the end or beginning of the sentence (defined as the encounter of a full stop, comma or question mark), whichever comes first. Preliminary experiments, not presented here due to space constraints, showed that the algorithm is quite robust to different neighborhood radii and the specific threshold was chosen as it is consistent with some prior work [Santos et al. 2009].

At this point, it is important to explain why this simple approach was chosen for modifier detection, rather than a more syntactically-correct approach of analyzing the text through a *parser* (e.g. Jiang and Liu [2010]) in order to extract the syntactic dependencies of the text. The reason is that, as mentioned above, the vast majority of informal textual communication contains significant spelling errors, making any such attempt very difficult and additionally seriously limiting the domain of applicability of the proposed solution. Some examples of such informal communication are provided later, in table II, where the datasets that are used in this study are presented.

If an *intensifier* or *diminisher* word is found then the original emotional value of the word is modified by the respective modifier score which is either added (in case of an intensifier) or subtracted (in case of a diminisher) to the absolute value of the term. This approach was adopted, instead of a more uniform ± 1 modification, because some modifiers are more intense than others, e.g. compare "fairly good" with "extremely good". For example, if "bad" has an initial value of -3 then "very bad" would be modified to -4. Similarly, "somewhat good" would be judged as +2, taking into consideration that "good" has an original value of +3.

If a *negation* term is found then the absolute value of the emotional term is decreased by 1 and its polarity is reversed. For example "not bad" would be +2. The intuition behind the reduction by one (instead of a simpler reversal of signs) is that although the polarity of a term is reversed with the usage of negation, the full original emotional weight of a term (such as "bad" in the above example) isn't fully transferred to the other class and thus the reduction by one. Simply put, one doesn't typically use "not bad" if one means "good".

Lastly, for the *capitalization* detection module, if a word, larger than two characters (in order to avoid false positives caused by normal article capitalization after a full stop), that is written fully in capital letters is detected within the neighborhood of an emotional word, including the actual emotional word, then the weight of the word is modified in the same manner as if an *intensifier* with a weight of 1 was detected. The *exclamation detection* module functions in the same manner. In contrast, *emoticons* are considered as explicit indicators of emotion [Derks et al. 2008] rather than modifiers and are assigned specific weights, i.e. +3 for positive emoticons and -3 for negative.

The score of a document on the C_{pos} and C_{neg} scales is the *maximum* positive and negative number produced respectively. As previously stated, for binary positive/negative prediction the class with the highest absolute value is considered dominant. A document is classified as objective if its scores are $\{+1, -1\}$. Algorithm 1 presents the full details of the classifier in pseudocode.

4. EXPERIMENTAL SETUP

We used three different datasets extracted from real-world social websites, in order to test the effectiveness of the proposed lexicon-based classifier.

The first data set was extracted from the social networking and microblogging website *Twitter*[4]. The dataset is comprised of two subsets. The first one (henceforth referred to as *Train*) was collected through the Twitter API, based on the existence of particular emoticons, which were used to provide an indication of the polarity of the text: positive for tweets that contain '`:)`' or '`:-)`' and negative for tweets that contain '`:(`' or '`:-(`'. Although this subset is at least two orders of magnitude larger than previous datasets, it is expected to be more "noisy" than human-annotated data, containing misclassified documents. It is expected nonetheless, and in fact the reason for the creation of datasets in such an automatic-fashion (e.g. [Go et al. ; Read 2005]) is that the abundance of training documents will override such errors and overall produce adequate supervised classifiers. The second subset (henceforth referred to as *Test*) was humanly annotated for objective, positive and negative emotion. More information on the dataset is provided by Pak and Paroubek [2010].

The second data set is from the social news website *Digg*[5], one of the most popular sites on the web where people share and discuss news and ideas. The site is very loosely administered and therefore any kind of language (including profanity) is allowed. The original data set spans the months February-April 2009 and contains about 1.6 million individual comments. The data set is described in detail by Paltoglou et al. [2010] and is freely available.

The last dataset is from the social website *MySpace* and it comprises of a sample of comments exchanged between friends in each other's public profile. The sample contains 40,000 profiles of members that joined on July 2007. All samples are filtered based on the country of origin of the user and only those that were based on UK or USA were kept. Thelwall and Wilkinson [2010] provide more information about the dataset.

A random subset of 1,000 comments was sampled from the last two datasets and

[4]`http://www.twitter.com`

[5]`http://www.digg.com`

Algorithm 1 Lexicon-Based classifier

INPUT: Affective Dictionary $LIWC$
INPUT: $AbbrevationList, NegationList, IntensifierList$
INPUT: $ExclamationList, DiminisherList, EmoticonList$
INPUT: Document $d = \{w_1, w_2, ...w_n\}$ to be classified
Initialize $C_{pos} \leftarrow +1$, $C_{neg} \leftarrow -1$, $PosInstances \leftarrow 0$, $NegInstances \leftarrow 0$
for all $w_i \in d$ **do**
  if $w_i \in AbbrevationList$ **then** $w_i \leftarrow FullForm(w_i)$
  end if
  if $w_i \in LIWC \cup EmoticonList$ **then**
    $temp_{w_i} \leftarrow EmotWeight(w_i)$
    $loffset \leftarrow lSentenceBound() \leq (i-5)?(i-5) : lSentenceBound()$
    $roffset \leftarrow rSentenceBound() \geq (i+5)?(i+5) : rSentenceBound()$
    for all w_k where $loffset \leq k \leq roffset$ **do**
      if $w_k \in NegationList$ **then**
        if $temp_{w_i} \leq 0$ **then**
          $temp_{w_i} \leftarrow -temp_{w_i} - 1$
        else
          $temp_{w_i} \leftarrow -temp_{w_i} + 1$
        end if
      end if
      if $w_k \in IntensifierList$ **then**
        if $temp_{w_i} \leq 0$ **then**
          $temp_{w_i} \leftarrow temp_{w_i} - IntenseWeight(w_k)$
        else
          $temp_{w_i} \leftarrow -temp_{w_i} + IntenseWeight(w_k)$
        end if
      end if
      if $w_k.length \geq 2$ AND $w_k = ALLCAPITALS$ **then**
        if $temp_{w_i} \leq 0$ **then**
          $temp_{w_i} \leftarrow temp_{w_i} - 1$
        else
          $temp_{w_i} \leftarrow -temp_{w_i} + 1$
        end if
      end if
      if $w_k \in DiminisherList$ **then**
        if $temp_{w_i} \leq 0$ **then**
          $temp_{w_i} \leftarrow temp_{w_i} + DiminishWeight(w_k)$
        else
          $temp_{w_i} \leftarrow -temp_{w_i} - DiminishWeight(w_k)$
        end if
      end if
      if $w_k \in ExclamationList$ **then**
        if $temp_{w_i} \leq 0$ **then**
          $temp_{w_i} \leftarrow temp_{w_i} - 1$
        else
          $temp_{w_i} \leftarrow temp_{w_i} + 1$

```
                end if
            end if
        end for
    end if
    if temp_{w_i} > 5 then
        temp_{w_i} ← 5
    end if
    if temp_{w_i} < −5 then
        temp_{w_i} ← −5
    end if
    if temp_{w_i} > 0 then
        PosInstances ← PosInstances + 1
    else
        NegInstances ← NegInstances + 1
    end if
    if temp_{w_i} > 0 AND temp_{w_i} > C_{pos} then
        C_{pos} ← temp_{w_i}
    end if
    if temp_{w_i} < 0 AND temp_{w_i} < C_{neg} then
        C_{neg} ← temp_{w_i}
    end if
end for
if C_{pos} = |C_{neg}| = 1 then return objective
else if C_{pos} > |C_{neg}| then return positive
else if C_{pos} < |C_{neg}| then return negative
else if C_{pos} = |C_{neg}| then
    if PosInstances > NegInstances then return positive
    else if PosInstances < NegInstances then return negative
    else if PosInstances = NegInstances then return objective
    end if
end if
```

given to 3 human assessors with the task of manually annotating their emotional content on two 5-point scales for positive and negative sentiment: `[no positive emotion or energy] +1,+2,...,+5 [very strong positive emotion]` and `[no negative emotion] -1,-2,...,-5 [very strong negative emotion]`. The reasons for the specific annotation scheme as well as more information about the annotation process and the inter-annotator agreement is provided by Paltoglou et al. [2010]. Note that the process was applied only to the last two datasets and that the *Twitter* corpus already contains human ternary annotations for positive, negative and neutral content.

We mapped the original two-dimensional 5-point scale human annotation to a binary scheme (objective vs. subjective and positive vs. negative) in the following manner:

—All the posts that have been rated by the majority of annotators (at least 2 out of 3) with scores -1 and +1 are considered "objective".

Table I. Number of documents per class for each data set used.

Data set	Number of Documents			Avg. Words per Document	Total Number of Documents
	Neutral	Positive	Negative		
Digg	144	107	221	32.91	472
MySpace	153	400	105	41.17	658
Twitter *Train*	-	232,442	151,955	19.69	384,397
Twitter (Test)	33	108	75	18.05	216

—All posts that have been rated by the majority of annotators (at least 2 out of 3) with a positive score equal or higher than +3 and a negative score of -1 or -2 are considered "positive".

—All posts that have been rated by the majority of annotators (at least 2 out of 3) with a negative score equal or lower than -3 and a positive score of +1 or +2 are considered "negative".

We use the union of positive and negative posts as "subjective". Although the above process results in a smaller subset of the original 1,000 posts per data set (see table I), it also guarantees a high inter-annotator agreement on a ternary scale and the remaining posts are much more definitive of their emotional content as some of the ambiguities of the original annotations are removed[6]. In the experiments that are presented we use this subset as the "gold standard" to train/test the machine-learning classifiers and evaluate the effectiveness of the lexicon-based classifier.

We used these three datasets because they provide typical examples of three inherently different types of online communication, i.e. an open discussion environment, a social networking site and a microblogging service, all of which are characteristic of the type of informal communication that takes place on the web. Additionally, although smaller in size compared to the typical datasets used in

[6]All the datasets are freely available upon request.

Table II. Examples of comments from all the datasets that were used in this study as annotated by human assessors, with the exception of the Twitter *Train* dataset. Because of space constraints, we only present some of the smaller comments.

Dataset	Category		
	Objective	Positive	Negative
Digg	uhhh... what?	He has a beautiful family photo! :)	nuke iran
MySpace	watz gud? c/b	i have free internet!!!! lol	All of you are OLD!!!!!!!!!
Twitter (Train)	-	im meeting up with one of my besties tonight! Cant wait!! :-) - GIRL TALK!!	Thinking I really hate packing, can anyone help please? :(
Twitter (Test)	testing Twitter API	@mikefish Fair enough. But i have the Kindle2	omg so bored & my tattoooos are so itchy!! help! aha =)

review-oriented sentiment analysis, e.g. [Pang et al. 2002], the aforementioned datasets are the largest, to our knowledge, that contain emotional annotations by human assessors of social media communication, therefore providing a "golden-standard" for classification. Table I presents some statistics about all three datasets and Table II presents some characteristic comments from each.

We used three machine-learning approaches in order to compare the effectiveness of the proposed lexicon-based classifier: Naive Bayes (NB) [Mccallum and Nigam 1998], Maximum Entropy (MaxEnt) [Nigam et al. 1999] and Support Vector Machines (SVM) [Joachims 1999; Chang and Lin 2001], using unigrams as features. All three have been used extensively in research and are considered state-of-the-art. Previous research in sentiment analysis [Pang et al. 2002] on review datasets has shown that binary features, which only capture the presence or absence of a term, often outperform term frequency features, where the weight of a feature is equal to the number of times that feature occurs in the document, but since no comparison has been conducted on the specific environment that we are examining here, we present results using both weighting schemes for the SVM and Naive Bayes classifiers[7], represented as SVM_{pr}, NB_{pr} and SVM_{tf}, NB_{tf} respectively. All machine-learning algorithms were implemented using the Weka Toolkit [Witten and Frank 1999][8].

We present results using 10-fold cross validation for the machine-learning approaches for the Digg, MySpace and Twitter-*Test* datasets for two classification tasks: objective vs. subjective and positive vs. negative. In addition, we use the Twitter-*Train* subset to train the classifiers on polarity classification and test them, in a hold-out evaluation, on the Twitter-*Test* subset. The lexicon classifier is tested on the complete datasets, since it doesn't require any reference corpus. Because of the fact that the datasets are unbalanced we also present results using two baselines: majority and random. In the former case, the class with the highest number of documents is always predicted and in the latter case, a class is selected at random.

As is typical with unbalanced datasets (e.g. [Li et al. 2005]) we present results based on the average value of the $F1$-score for both categories [Manning et al. 2008] to quantify classification quality. Category-based precision ($Pr.$), recall ($R.$) and F1 are also reported for completeness reasons. The $F1$ for class i is defined as:

$$F1_i = \frac{2Pr_iR_i}{R_i + Pr_i}$$

where Pr_i and R_i are the precision and recall that the classifier attains for class i respectively, defined as:

$$Pr_i = \frac{tp}{tp + fp}, \quad R_i = \frac{tp}{tp + fn}$$

where tp is the number of documents correctly classified as belonging to class i ("true positive"), fp is the number of documents falsely classified as belonging to class i ("false positive") and fn is the number of documents falsely classified as not

[7]The Maximum Entropy classifier uses only binary features.
[8]All the Weka *.arff* files for all datasets are available upon request.

Table III. Polarity classification on the Digg dataset (%). Notation is explained in section 4.

	Positive			Negative			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	0.0	0.0	0.0	67.4	100	80.5	40.3
Random	32.6	50.0	39.4	67.4	50.0	57.4	48.4
Lexicon	60.0	**89.7**	**71.9**	**93.3**	70.8	80.5	**76.2**
SVM_{pr}	60.8	68.2	64.3	83.7	78.7	81.1	72.7
SVM_{tf}	59.8	68.2	63.8	83.5	77.8	80.6	72.2
NB_{pr}	53.7	73.8	62.2	84.5	69.2	76.1	69.2
NB_{tf}	**69.5**	38.8	49.4	75.5	**91.9**	**82.9**	66.2
MaxEnt	56.1	69.2	61.9	83.2	73.8	78.2	70.1

belonging to class i ("false negative"). The final average $F1$ measure is calculated as $F1 = \frac{1}{|i|} \sum_i F1_i$.

5. RESULTS - DISCUSSION

The results for polarity classification are shown in tables III to V and for subjectivity detection are shown in tables VI to VIII.

5.1 Polarity classification

The results from the polarity classification task on the Digg dataset are presented in table III. The majority and random approaches provide a measure of baseline results. As it can be seen their average F1 value ranges from 40.3% for the former to 48.4% for the latter. The machine-learning approaches perform, as expected, much better with the two SVM approaches overall outperforming the Naive Bayes and the Maximum Entropy classifier and also the latter outperforming the former. Additionally, presence-based features outperform frequency-based features, a result which is in accordance with previous research [Pang et al. 2002] providing an potential indication for the existence of some "universal" concepts for sentiment analysis, whether it is applied in product reviews or social communication. The lexicon-based classifier noteworthily outperforms the best of the supervised approaches (76.2% vs. 72.7% average F1), providing a first indication of its potential. Taking into consideration the diversity of topics discussed on Digg and the liberal usage of language at the website, the results can be considered very encouraging.

Results from sentiment analysis on the MySpace dataset are provided in table IV. In this setting, the Naive Bayes classifier using binary features performs unexpect-

Table IV. Polarity classification on the MySpace dataset (%). Notation is explained in section 4.

	Positive			Negative			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	79.2	100	88.4	0.0	0.0	0.0	44.2
Random	79.2	50.0	61.3	20.8	50.0	29.3	45.3
Lexicon	**93.2**	92.5	**92.9**	67.1	69.5	**68.3**	**80.6**
SVM_{pr}	87.4	93.5	90.3	66.2	48.6	56.0	73.2
SVM_{tf}	86.2	90.8	88.4	56.0	**79.9**	49.7	69.1
NB_{pr}	88.4	89.3	88.8	57.4	55.2	56.3	72.6
NB_{tf}	83.9	**98.0**	90.4	**78.9**	28.6	42.0	66.2
MaxEnt	83.9	91.0	87.3	49.3	33.3	39.8	63.6

Table V. Polarity classification on the Twitter dataset (%). 10-fold cross validation is done on the Twitter *Test* subset and hold-out validation uses the Twitter *Train* subset for training and Twitter *Test* for testing. Notation is explained in section 4.

	Positive			Negative			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	59.0	100	74.2	0.0	0.0	0.0	37.1
Random	59.0	50.0	54.1	41.0	50.0	45.0	49.6
Lexicon	**86.1**	**93.5**	**89.7**	**89.3**	**78.1**	**83.3**	**86.5**
10-fold cross validation							
SVM_{pr}	72.9	79.6	76.1	66.2	57.3	61.4	68.8
SVM_{tf}	77.0	80.6	78.7	70.0	65.3	67.6	73.2
NB_{pr}	77.8	77.8	77.8	68.0	68.0	68.0	72.9
NB_{tf}	80.2	78.7	79.4	70.1	72.0	71.1	75.3
MaxEnt	80.0	74.1	76.9	66.3	73.3	69.6	73.3
Hold-out validation							
SVM_{pr}	75.7	77.8	76.7	66.7	64.0	65.3	71.0
SVM_{tf}	74.6	78.7	76.6	66.7	61.3	63.9	70.3
NB_{pr}	78.6	81.5	80.0	71.8	68.0	69.9	75.0
NB_{tf}	78.4	84.3	81.3	74.6	66.7	70.4	76.9
MaxEnt	80.9	90.7	85.5	83.8	69.3	75.9	80.7

edly well, even though still lower than the SVM_{pr} classifier. Closer examination of the class-based results of the NB classifier nonetheless shows that the result is somewhat "artificial" and doesn't reflect true effectiveness. Indeed, the result is mainly caused by the fact that the dataset is heavily unbalanced towards positive documents[9] and the well-documented preference of the Naive Bayes classifier for the more popular class. The lexicon-based classifier again manages to outperform every other approach, both in average F1 (80.6% vs. 73.2%) and even in class-based F1 values, 92.9% vs. 90.4% for the positive class and 68.3% vs. 56.3% for the negative (comparisons are always with the best performing supervised classifier at the particular metric). The results demonstrate one of the key advantages of using an unsupervised classifier in social-networking environments where gathering training data isn't only very time-consuming and requires human-labor (as in this case) but can often result in heavily unbalanced data, which, produce models that may have undesired biases. In comparison, the lexicon-based classifier can be applied "off-the-shelf" without any modification or training and still produce good results.

Table V presents results from the Twitter dataset and are of particular interest, because of the existence of the significantly larger *Train* subset. In the 10-fold cross validation setting, the results are similar to the previous two settings, with the newly proposed solution managing to outperform all machine-learning approaches, by a considerable margin in almost every metric, e.g. 86.5% vs. 75.3% for average F1.

We focus on the performance of the lexicon-based classifier against the machine-learning approaches that have been trained on the more voluminous *Train* subset. As expected therefore, most of the algorithms, with the exception of the SVM classifier using term frequency features, perform better, taking advantage of the abundance of training documents. It must be noted that the performance of the machine-learning approaches in this setting isn't greatly improved over the results

[9]There are four times as many positive than negative documents in this dataset, see table I.

Table VI. Subjectivity detection on the Digg dataset. Notation is explained in section 4.

	Objective			Subjective			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	0.0	0.0	0.0	69.4	100	82.0	41.0
Random	30.5	50.0	37.9	69.4	50.0	58.2	48.0
Lexicon	63.6	76.4	**69.4**	**88.6**	80.8	**84.5**	**77.0**
SVM_{pr}	58.2	66.7	62.1	84.4	79.0	81.6	71.9
SVM_{tf}	58.8	65.3	61.8	84.0	79.9	81.9	71.9
NB_{pr}	54.4	**77.8**	64.0	88.0	71.3	78.8	66.6
NB_{tf}	**71.5**	36.8	48.6	77.1	**93.6**	**84.5**	71.4
MaxEnt	42.0	60.4	49.6	78.5	63.4	70.2	59.9

produced in the previous setting, despite the existence of an order of magnitude more training data, demonstrating that applying sentiment analysis in informal textual communication is a particularly challenging problem. Nonetheless, the proposed unsupervised classifier is still able to outperform all approaches obtaining an average F1 value of 86.5% vs. the best performing supervised solution of 80.7% ($MaxEnt$). The specific results are of particular importance because they demonstrate that even in environments where massive training data can be collected through an automatic or semi-automatic fashion, the lexicon-based classifier is still able to outperform the produced supervised models.

Overall, the results from the polarity classification task show that the new classifier offers a very robust performance and is able to outperform machine-learning classifiers in most cases, even in settings where there is a significant number of training documents. This result is very important as it clearly demonstrates the robustness of the approach to different settings and social environments. The significant advantage of the method is that it requires no training and can be applied without any modification, therefore it could easily and readily be applied to other similar environments which are not tested here, such as Facebook, blogs and forums, offering overall a very robust solution.

5.2 Subjectivity classification

We change our focus from polarity classification to subjectivity classification, i.e. the detection of whether segments of text contain opinions or are objective. As discussed in section 4, we have used the union of positive and negative documents as subjective (see table I).

Table VI presents the first results of this task on the Digg dataset. Similarly to the polarity classification task, the lexicon-based classifier is able to outperform machine-learning approaches: 77.0% vs. 71.9% for the best performing supervised approach (SVM_{pr}). Taking into consideration that the lexicon-based classifier also demonstrated very good performance in the previous task in the same dataset, it can be concluded that the algorithm provides a very good solution to the general problem of initially detecting and subsequently classifying opinions in such open discussion systems.

Results from the subjectivity detection task on the MySpace dataset are provided in table VII. Again, the lexicon-based classifier is able to outperform, although only slightly, supervised approaches: 79.7% against 79.2% for the best performing

supervised approach (SVM_{tf}). The results indicate that the proposed solution isn't only able to effectively detect subjectivity in open discussion websites, such as Digg, but is also effective in social networking sites, such as MySpace and therefore potentially Facebook, Flickr, etc.

Lastly, we present results from the Twitter dataset for the task of subjectivity detection at table VIII. This is the only setting that the machine-learning approaches outperform the lexicon-based classifier, which attains an average F1 of 70.9%. The SVM_{pr} classifier performs best in this setting (75.3%), followed by the SVM_{tf} (72.1%) and then by the Naive Bayes classifiers using binary (71.7%) features. Peculiarly enough, the lexicon classifier still outperforms the Maximum Entropy (63.9%), which was the best performing machine-learning classifier in the polarity classification task in the Twitter dataset (table V). Careful analysis of the class-based metrics in this environment reveal that the lexicon-based approach attains the best recall for the objective class and is therefore able to correctly identify most of the objective documents in this dataset while at the same time also attaining the best precision for subjective tweets, overall demonstrating that even when the classifier doesn't offer the best effectiveness over machine-learning approaches, it still manages to perform very successfully for some specific subtasks (i.e. a high-precision detection task for subjective documents).

Overall, the results from the subjectivity detection task show that the lexicon-based classifier proposed in this paper is able to perform very adequately in the majority of environments and overall offers a very reliable solution. The fact that the classifier was tested in three inherently different types of online environments, i.e. an open discussion environment, a social networking site and a microblogging service, without any modifications or training, provides a clear demonstration of its potential.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of sentiment analysis in social networking media, such as MySpace, Twitter, Digg, forums, blogs, etc. We argued that this area of application provides unique challenges, not addressed in typical review-focused sentiment analysis environments.

We proposed an intuitive, unsupervised, lexicon-based algorithm which estimates the level of emotional strength in text in order to make a final prediction. Our proposed solution is applicable to two complimentary tasks: subjectivity detection and

Table VII. Subjectivity detection on the MySpace dataset. Notation is explained in section 4.

	Objective			Subjective			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	0.0	0.0	0.0	76.7	100	87.0	43.4
Random	23.2	50.0	31.7	76.7	50.0	60.6	46.1
Lexicon	68.2	69.9	**69.0**	90.8	90.1	**90.5**	**79.7**
SVM_{pr}	65.2	68.6	66.9	90.3	88.9	89.6	78.6
SVM_{tf}	67.1	69.3	68.2	90.6	89.7	90.1	79.2
NB_{pr}	59.5	**73.9**	65.9	**91.5**	84.8	88.0	77.0
NB_{tf}	**76.8**	28.1	41.1	81.7	**97.4**	88.9	65.0
MaxEnt	50.0	75.2	60.1	91.1	77.2	83.6	71.9

Table VIII. Subjectivity detection on the Twitter *Test* dataset (%). Notation is explained in section 4.

	Objective			Subjective			Avg.
	Pr.	R.	F1	Pr.	R.	F1	F1
Majority	0.0	0.0	0.0	84.7	100	91.7	45.8
Random	15.2	50.0	23.4	84.7	50.0	62.9	43.1
Lexicon	44.7	**63.6**	52.5	**92.9**	85.8	89.2	70.9
SVM_{pr}	**69.6**	48.5	**57.1**	91.2	**96.2**	**93.6**	**75.3**
SVM_{tf}	60.0	45.5	51.7	90.6	94.5	92.5	72.1
NB_{pr}	48.7	57.6	52.8	92.1	89.1	90.6	71.7
NB_{tf}	47.4	54.5	50.7	91.6	89.1	90.3	70.5
MaxEnt	34.7	51.5	41.5	90.4	82.5	86.3	63.9

polarity classification, overall providing a comprehensive solution to the problem of sentiment analysis of informal communication on the web. The advantages of the approach is that it requires no training and thus can be readily applied into a wide selection of environments.

We used three different real-world, humanly annotated datasets to compare the effectiveness of the classifier against state-of-the-art machine learning approaches. All datasets were extracted from popular websites, are publicly available and are indicative of the diverse set of environments that are available to users today: an open news-sharing and discussion forum, a social networking website and a microblogging service. Although, naturally, we haven't exhausted the list of potential environments, we believe that the specific datasets provide an indicative list of relevant settings.

In the polarity classification task, the newly proposed classifier was able to outperform machine learning approaches in the majority of experiments, even in settings where significant training data was available. Similarly, in the subjectivity detection task the lexicon-based approach was able to perform very well in the majority of environments, outperforming other solutions in most cases. Overall, the experimental results demonstrated that the proposed solution, although simple in its conception offers a very robust and reliable solution.

In the future, we plan to incorporate the ANEW list of words [Bradley and Lang 1999], which readily provides emotional weights for tokens on a 1-9 scale, to the list of utilized lexicons used by the algorithm. Additionally, we plan to expand the emotional prediction capabilities of the algorithm to all three dimensions that the ANEW provides, i.e. arousal and dominance in addition to valence, thus providing a more emotionally comprehensive analysis of textual communication.

In order to further improve the performance of the lexicon-based approach we plan to incorporate machine learning techniques to optimize the emotional weights of tokens and modifiers. The aim of this approach would to make the algorithm more adaptable and easily configurable to different and novel environments.

Despite the fact that in this work we utilized a static emotional lexicon that can be easily manually expanded if new affective words or expressions become popular in social media, we intend to explore methods of automatically expanding it without human intervention. Such a process could function in a boot-strapping fashion, using the already existing lexicon to classify documents and extracting new words that have a significant discriminative power, based on some feature selection

criteria, such as information gain, for addition to the emotional lexicon.

Lastly, we also aim to experiment with non-English text, by using different, language-dependent emotional lexicons and translating the already available ones. Our goal is to extend the applicability of the proposed solution to other languages, for which training data is especially difficult to come by.

7. ACKNOWLEDGMENTS

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions Project (Contract 231323).

REFERENCES

Baccianella, S., Esuli, A., and Fabrizio, S. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC'10* (19-21).

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. Feature selection for ordinal regression. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, New York, NY, USA, 1748–1754.

Blitzer, J., Dredze, M., and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL '07*. 440–447.

Bradley, M. and Lang, P. 1999. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Tech. rep., Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Brooke, J., Tofiloski, M., and Taboada, M. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of ICRA-NLP '09*.

Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chung, C. and Pennebaker, J. 2007. The psychological function of function words. *Social communication: Frontiers of social psychology*, 343–359.

Cornelius, R. R. 1996. *The science of emotion*. Prentice Hall.

Dave, K., Lawrence, S., and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of WWW '03*. 519–528.

Derks, D., Bos, A. E. R., and Von Grumbkow, J. 2008. Emoticons and online message interpretation. *Soc. Sci. Comput. Rev. 26,* 3, 379–388.

Devitt, A. and Ahmad, K. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of ACL '07*. 984–991.

Ding, X., Liu, B., and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of WSDM '08*. 231–240.

Fox, E. 2008. *Emotion Science*. Palgrave Macmillan.

Go, A., Huang, L., and Bhayani, R. Tech. rep., Stanford Natural Language Processing Group, 2009.

Harvey, M. 2010. Facebook ousts google in us popularity. March 17, 2010, The Sunday Times, last accessed July 05, 2010, `http://technology.timeson-line.co.uk/tol/news/tech\_and\_web/the\_web/article7064973.ece`.

Hassan, A. and Radev, D. R. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 395–403.

Jiang, W. and Liu, Q. 2010. Dependency parsing and projection based on word-pair classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 12–20.

Joachims, T. 1999. Making large-scale support vector machine learning practical. 169–184.

Kennedy, A. and Inkpen, D. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence 22,* 2, 110–125.

Li, Y., Bontcheva, K., and Cunningham, H. 2005. Using uneven margins svm and perceptron for information extraction. In *Proceedings of CONLL '05*. 72–79.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of CoNLL '06*.

Manning, C. D., Raghavan, P., and Schtze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Mccallum, A. and Nigam, K. 1998. A comparison of event models for naive bayes text classification.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *1st Workshop on Stylistic Analysis Of Text For Information Access*.

Mullen, T. and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP 2004*. 412–418.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. 2007. Textual affect sensing for sociable and expressive online communication. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*. ACII '07. Springer-Verlag, Berlin, Heidelberg, 218–229.

Nigam, K., Lafferty, J., and Mccallum, A. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*. 61–67.

Osgood, C. E. 1967. *The measurement of meaning / [by] [Charles E. Osgood, George J. Suci [and] Percy H. Tannenbaum]*, 2nd ed. ed. University of Illinois Press, Urbana :.

Pak, A. and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC'10*.

Paltoglou, G., Thelwall, M., and Buckely, K. 2010. Online textual communcation annotated with grades of emotion strength. In *Proceedings of the Third International Workshop on EMOTION (satellite of LREC): Corpora for research on emotion and affect*. 25–31.

Pang, B. and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL'04*. 271–278.

Pang, B. and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*.

Pennebaker J., F. M. and R., B. 2001. *Linguistic Inquiry and Word Count: LIWC*, 2 ed. Erlbaum Publishers.

Qiu, L., Zhang, W., Hu, C., and Zhao, K. 2009. SELC: a self-supervised model for sentiment classification. In *Proceeding CIKM '09*. 929–936.

Quirk, R. 1985. *A comprehensive grammar of the English language / Randolph Quirk ... [et. al.] ; index by David Crystal*. Longman, London ; New York :.

Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop on - ACL '05*. Number June. Association for Computational Linguistics, Morristown, NJ, USA, 43.

Santos, R. L. T., He, B., Macdonald, C., and Ounis, I. 2009. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. ECIR '09. Springer-Verlag, Berlin, Heidelberg, 325–336.

Schimmack, U. 2001. Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive? *Cognition and Emotion 15,* 1 (January), 81–97.

Slatcher, R., Chung, C., Pennebaker, J., and Stone, L. 2007. Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality 41,* 1, 63–75.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Takamura, H., Inui, T., and Okumura, M. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational*

Linguistics. ACL '05. Association for Computational Linguistics, Morristown, NJ, USA, 133–140.

Thelwall, M. 2008. Social networks, gender, and friending: An analysis of myspace member profiles. *J. Am. Soc. Inf. Sci. Technol. 59,* 8, 1321–1330.

Thelwall, M. 2009. Myspace comments. *Online Information Review 33,* 1, 58–76.

Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology (JASIST)*. in press.

Thelwall, M. and Wilkinson, D. 2010. Public dialogs in social network sites: What is their purpose? *J. Am. Soc. Inf. Sci. Technol. 61,* 2, 392–404.

Thomas, M., Pang, B., and Lee, L. 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP '06*. 327–335.

Turney, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*. 417–424.

Turney, P. D. and Littman, M. L. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *CoRR cs.LG/0212012*.

Whitelaw, C., Garg, N., and Argamon, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of CIKM '05*. 625–631.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, Morristown, NJ, USA, 34–35.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005a. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 347–354.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *In Proceedings of HLT/EMNLP 2005*.

Wilson, T., Wiebe, J., and Hoffmann, P. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Linguist. 35*, 399–433.

Witten, I. H. and Frank, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*, 1st ed. Morgan Kaufmann.

Wright, A. 2009. Mining the web for feelings, not facts. August 23, NY Times, last accessed October 2, 2009, `http://http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?\_r=1`.

Yessenalina, A., Choi, Y., and Cardie, C. 2010. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, 336–341.

Zaidan, O., Eisner, J., and Piatko, C. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. *Proceedings of NAACL HLT*, 260–267.